Chromatic PAC-Bayes Bounds for Non-IID Data

Liva Ralaivola, Marie Szafranski, Guillaume Stempfel

Chromatic PAC-Bayes Bounds for Non-IID Data

Liva Ralaivola, Marie Szafranski, Guillaume Stempfel
Laboratoire d'Informatique Fondamentale de Marseille
Aix-Marseille Universités
39, rue F. Joliot Curie, F-13013 Marseille, France
firstname.lastname@lif.univ-mrs.fr

Abstract

PAC-Bayes bounds are among the most accurate generalization bounds for classifiers learned with IID data, and it is particularly so for margin classifiers. However, there are many practical cases where the training data show some dependencies and where the traditional IID assumption does not apply. Stating generalization bounds for such frameworks is therefore of the utmost interest, both from theoretical and practical standpoints. In this work, we propose the first – to the best of our knowledge – PAC-Bayes generalization bounds for classifiers trained on data exhibiting interdependencies. The approach undertaken to establish our results is based on the decomposition of a so-called dependency graph that encodes the dependencies within the data, in sets of independent data, through the tool of graph fractional covers. Our bounds are very general, since being able to find an upper bound on the (fractional) chromatic number of the dependency graph is sufficient to get new PAC-Bayes bounds for specific settings. We show how our results can be used to derive bounds for bipartite ranking and windowed prediction on sequential data.

1 Introduction

Over the past decade, there has been much progress in the field of generalization bounds for classifiers. PAC-Bayes bounds, introduced in (McAllester, 1999), and refined in, e.g., (Seeger, 2002; Langford, 2005), are among the most appealing advances. Their possible tightness, as shown in (Ambroladze et al., 2007), make them a possible route to perform model selection. They can also be viewed as theoretical tools to motivate new learning procedures.

Unfortunately, so far, PAC-Bayes bounds have only tackled the case of classifiers trained from *independently and identically distributed* (IID) data. Yet, being able to learn from non-IID data while having strong theoretical guarantees is an actual problem in a number of real world applications such as, e.g., k-partite ranking or classification from sequential data. Here, we propose the first PAC-Bayes bounds for classifiers trained on non-IID data; they are a direct generalization of the IID PAC-Bayes bound and they are general enough to provide a principled way to establish generalization bounds for a number of non-IID settings. To derive these new bounds, we only make use of standard and simple tools of probability theory, convexity properties of adequate functions, and we exploit the notion of fractional covers of graphs. This tool from graph theory has already been used for deriving concentration results for non independent data in (Janson, 2004) (see also references of work making use of such decompositions therein) and for providing generalization bounds based on the so-called fractional Rademacher complexity by (Usunier et al., 2006).

The paper is organized as follows. Section 2 first recalls the standard IID PAC-Bayes bound, introduces the notion of fractional covers of graphs and then states the new *chromatic* PAC-*Bayes bounds*, called so, because they rely on the fractional chromatic number of a particular graph, namely the *dependency graph* of the data at hand. Section 3 is devoted to the proof of our main theorem. In Section 4, we provide specific versions of one of our bounds for the case of IID data, showing that it is a direct generalization of the standard bounds, for the case of bipartite ranking and for windowed prediction on sequential data.

Appearing in Proceedings of the 12^{th} International Confe-rence on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5.

2 PAC Bayes Bounds and Fractional Covers

2.1 IID PAC-Bayes Bound

Let us introduce some notation that will hold from here on. We only consider the problem of binary classification over the *input space* $\mathcal{X}$ and we denote $\mathcal{Z}$ the product space $\mathcal{X} \times \mathcal{Y}$, with $\mathcal{Y} = \{-1, +1\}$. $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is a family of classifiers from $\mathcal{X}$. D is a probability distribution defined on $\mathcal{Z}$ and $\mathbf{D}_m$ the distribution of an m-sample; for instance, $\mathbf{D}_m = \otimes_{i=1}^m D = D^m$ is the distribution of an IID sample $\mathbf{Z} = \{Z_i\}_{i=1}^m$ of size m, with each Z_i distributed according to D. P and Q are distributions over $\mathcal{H}$.

The usual PAC-Bayes bound, can be stated as follows (McAllester, 1999; Seeger, 2002).

Theorem 1 (IID PAC-Bayes Bound)**.** *$\forall m$, $\forall D$, $\forall \mathcal{H}$, $\forall \delta \in (0, 1]$, $\forall P$, with probability at least $1 - \delta$ over the random draw of $\mathbf{Z} \sim \mathbf{D}_m = D^m$, the following holds:*

$$\forall Q, kl(\hat{e}_Q || e_Q) \leq \frac{1}{m} \left[KL(Q||P) + \ln \frac{m+1}{\delta} \right]. \tag{1}$$

This theorem actually provides a generalization error bound for the so-called Gibbs classifier g_Q: given a distribution Q over $\mathcal{H}$, this stochastic classifier predicts a class for an input $\mathbf{x} \in \mathcal{X}$ by first drawing a hypothesis h according to Q and then outputting $h(\mathbf{x})$. In the theorem, $\hat{e}_Q$ is the empirical error of g_Q on an IID sample $\mathbf{Z}$ of size m and e_Q is its true error:

$$\begin{aligned} \hat{e}_Q &= \mathbb{E}_{h \sim Q} \frac{1}{m} \sum_{i=1}^m r(h, Z_i) = \mathbb{E}_{h \sim Q} \hat{R}(h, \mathbf{Z}) \\ e_Q &= \mathbb{E}_{\mathbf{Z} \sim \mathbf{D}_m} \hat{e}_Q = \mathbb{E}_{\substack{Z \sim D \\ h \sim Q}} r(h, Z) = \mathbb{E}_{h \sim Q} R(h), \end{aligned} \tag{2}$$

where, for $Z = (X, Y)$, $r(h, Z) = \mathbb{I}_{h(X) \neq Y}$ and where we have used the fact that $\mathbf{Z}$ is an (independently) identically distributed sample. $\mathrm{kl}(q||p)$ is the Kullback-Leibler divergence between the Bernoulli distributions with probabilities of success q and p, and $\mathrm{KL}(Q||P)$ is the Kullback-Leibler divergence between Q and P:

$$\begin{aligned} \mathrm{kl}(q||p) &= q \ln \frac{q}{p} + (1-q) \ln \frac{1-q}{1-p} \\ \mathrm{KL}(Q||P) &= \mathbb{E}_{h \sim Q} \ln \frac{Q(h)}{P(h)}. \end{aligned}$$

Throughout the paper, we make the assumption that the posteriors that are used are absolutely continuous with respect to their corresponding priors.

We note that even if the present bound does apply to the risk e_Q of the stochastic classifier g_Q, a straightforward argument gives that, if b_Q is the (deterministic) Bayes classifier such that $b_Q(x) = \mathrm{sign}(\mathbb{E}_{h \sim Q} h(x))$, then $R(b_Q) \leq 2e_Q$ (Langford & Shawe-taylor, 2002).

The problem we are interested in in the present work is that of generalizing Theorem 1 to the situation where there may exist probabilistic dependencies between the elements Z_i of $\mathbf{Z} = \{Z_i\}_{i=1}^m$ but while, at the same time, the marginal distributions of the Z_i's are identical. In other words, we provide PAC-Bayes bounds for classifiers trained on *identically but not independently distributed data*. These results rely on properties of a dependency graph that is built according to the dependencies within $\mathbf{Z}$. Before stating our new bounds, we thus introduce the concepts of graph theory that will play a role in their statements.

2.2 Dependency Graphs and Fractional Covers

Definition 1 (Dependency Graph)**.** Let $\mathbf{Z} = \{Z_i\}_{i=1}^m$ be a set of random variables taking values in some space $\mathcal{Z}$. The *dependency graph* $\Gamma(\mathbf{Z})$ of $\mathbf{Z}$ is such that: the set of vertices of $\Gamma(\mathbf{Z})$ is $\{1, \ldots, m\}$ and there is an edge between i and j *if and only if* Z_i and Z_j are not independent (in the probabilistic sense).

Definition 2 (Fractional Covers, (Schreinerman & Ullman, 1997))**.** Let $\Gamma = (V, E)$ be an undirected graph, with $V = \{1, \ldots, m\}$.

- $C \subseteq V$ is *independent* if the vertices in C are independent (no two vertices in C are connected).
- $\mathbf{C} = \{C_j\}_{j=1}^n$, with $C_j \subseteq V$, is a *proper cover* of V if each C_j is independent and $\bigcup_{j=1}^n C_j = V$. The size of $\mathbf{C}$ is n.
- $\mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n$, with $C_j \subseteq V$ and $\omega_j \in [0, 1]$, is a *proper exact fractional cover* of V if each C_j is independent and $\forall i \in V$, $\sum_{i=1}^n \omega_j \mathbb{I}_{i \in C_j} = 1$; $\omega(\mathbf{C}) = \sum_{j=1}^n w_i$ is the *chromatic weight* of $\mathbf{C}$.
- $\chi(\Gamma)$ ($\chi^*(\Gamma)$) is the minimum size (weight) over all proper exact (fractional) covers of Γ: it is the *(fractional) chromatic number* of Γ.

The problem of computing the (fractional) chromatic number of a graph is known to be NP-hard (Schreinerman & Ullman, 1997). However, it turns out that for some particular graphs as those that come from the settings we study in section 4, this number can be evaluated precisely. The following properties hold (Schreinerman & Ullman, 1997):

Property 1. *Let $\Gamma = (V, E)$ be a graph. Let $c(\Gamma)$ be the* clique number *of Γ, i.e. the order of the largest clique in Γ. Let $\Delta(\Gamma)$ be the maximum degree of a vertex in Γ.*

We have the following inequalities:

$$1 \leq c(\Gamma) \leq \chi^*(\Gamma) \leq \chi(\Gamma) \leq \Delta(\Gamma) + 1.$$

In addition, $1 = c(\Gamma) = \chi^*(\Gamma) = \chi(\Gamma) = \Delta(\Gamma) + 1$ if and only if Γ *is totally disconnected.*

Remark 1. A cover can be thought of a fractional cover with every w_i being equal to 1. Hence, all the results that we state for fractional covers apply to the case of covers.

Remark 2. If $\mathbf{Z} = \{Z_i\}_{i=1}^m$ is a set of random variables over $\mathcal{Z}$ then a (fractional) proper cover of $\Gamma(\mathbf{Z})$, splits $\mathbf{Z}$ into subsets of independent random variables. This is a crucial feature to establish the results of the present paper.

In addition, we can see $\chi^*(\Gamma(\mathbf{Z}))$ and $\chi(\Gamma(\mathbf{Z}))$ as measures of the amount of dependencies within $\mathbf{Z}$.

The following lemma, also taken from (Janson, 2004), Lemma 3.1, will be very useful in the following.

Lemma 1. *If* $\mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n$ *is an exact fractional cover of* $\Gamma = (V, E)$*, with* $V = \{1, \ldots, m\}$*, then*

$$\forall \mathbf{t} \in \mathbb{R}^m,\ \sum_{i=1}^m t_i = \sum_{j=1}^n \omega_j \sum_{k \in C_j} t_k.$$

In particular $m = \sum_{j=1}^n |C_j|$.

2.3 Chromatic PAC-Bayes Bounds

In this subsection, we provide new PAC-Bayes bounds that apply for classifiers trained from samples $\mathbf{Z}$ according to distributions $\mathbf{D}_m$ where dependencies exist. We assume that those dependencies are fully determined by $\mathbf{D}_m$ and we can define the dependency graph $\Gamma(\mathbf{D}_m)$ of $\mathbf{D}_m$ to be $\Gamma(\mathbf{D}_m) = \Gamma(\mathbf{Z})$. As stated before, we make the assumption that the marginal distributions of $\mathbf{D}_m$ along each coordinate are equal to some distribution D.

We consider the following additional notation. $\textsc{Pefc}(\mathbf{D}_m)$ is the set of proper exact fractional covers of $\Gamma(\mathbf{D}_m)$. Given a cover $\mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n \in \textsc{Pefc}(\mathbf{D}_m)$, $\mathbf{Z}^{(j)} = \{Z_k\}_{k \in C_j}$ and $\mathbf{D}_m^{(j)}$ is the distribution of $\mathbf{Z}^{(j)}$, it is therefore equal to $D^{|C_j|}$; $\boldsymbol{\alpha} \in \mathbb{R}^n$ is the vector of coefficients $\alpha_j = \omega_j/\omega(\mathbf{C})$ and $\boldsymbol{\pi} \in \mathbb{R}^n$ is the vector of coefficients $\pi_j = \omega_j |C_j|/m$. $\mathbf{P}_n$ and $\mathbf{Q}_n$ are distributions over $\mathcal{H}^n$, P_n^j and Q_n^j are the marginal distributions of $\mathbf{P}_n$ and $\mathbf{Q}_n$ with respect to the jth coordinate, respectively; $\mathbf{h} = (h_1, \ldots, h_n)$ is an element of $\mathcal{H}^n$.

We can now state our main results.

Theorem 2 (Chromatic PAC-Bayes Bound (I))**.** $\forall m, \forall \mathbf{D}_m$, $\forall \mathcal{H}, \forall \delta \in (0,1]$, $\forall \mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n \in \textsc{Pefc}(\mathbf{D}_m)$, $\forall \mathbf{P}_n$, *with probability at least* $1-\delta$ *over the random draw of* $\mathbf{Z} \sim \mathbf{D}_m$*, the following holds:*

$$\forall \mathbf{Q}_n, kl(\bar{e}_{\mathbf{Q}_n} || e_{\mathbf{Q}_n}) \leq \frac{\omega}{m}\left[\sum_{j=1}^n \alpha_j KL(Q_n^j || P_n^j) + \ln \frac{m+\omega}{\delta\omega}\right], \tag{3}$$

where ω *stands for* $\omega(\mathbf{C})$ *and*

$$\begin{aligned}
\bar{e}_{\mathbf{Q}_n} &= \mathbb{E}_{\mathbf{h} \sim \mathbf{Q}_n} \frac{1}{m} \sum_{j=1}^n \omega_j \sum_{k \in C_j} r(h_j, Z_k) \\
&= \frac{1}{m} \sum_{j=1}^n \omega_j |C_j| \mathbb{E}_{h \sim Q_n^j} \frac{1}{|C_j|} \sum_{k \in C_j} r(h, Z_k) \\
&= \sum_{j=1}^n \pi_j \mathbb{E}_{h \sim Q_n^j} \hat{R}(h, \mathbf{Z}^{(j)}).
\end{aligned}$$

As usual, $e_{\mathbf{Q}_n} = \mathbb{E}_{\mathbf{Z} \sim \mathbf{D}_m} \bar{e}_{\mathbf{Q}_n}$.

The proof of this theorem is deferred to Section 3.

Remark 3. The empirical error $\bar{e}_{\mathbf{Q}_n}$ considered in this theorem is a weighted average of the empirical errors on $\mathbf{Z}^{(j)}$ of Gibbs classifiers with respective distributions Q_n^j.

The following proposition characterizes $\mathbb{E}_{\mathbf{Z} \sim \mathbf{D}_m} \bar{e}_{\mathbf{Q}_n}$.

Proposition 1. $\forall m$*,* $\forall \mathbf{D}_m$*,* $\forall \mathcal{H}$*,* $\forall \mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n \in \textsc{Pefc}(\mathbf{D}_m)$*,* $\forall \mathbf{Q}_n$*,*
$e_{\mathbf{Q}_n} = \mathbb{E}_{\mathbf{Z} \sim \mathbf{D}_m} \bar{e}_{\mathbf{Q}_n}$ *is the error of the Gibbs classifier based on the mixture of distributions* $Q^{\boldsymbol{\pi}} = \sum_{j=1}^n \pi_j Q_n^j$ *over* $\mathcal{H}$*.*

Proof. From Definition 2, $\pi_j \geq 0$ and, according to Lemma 1, $\sum_{j=1}^n \pi_j = \frac{1}{m} \sum_{j=1}^n \omega_j |C_j| = 1$.

Then,

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z} \sim \mathbf{D}_m} \bar{e}_{\mathbf{Q}_n} &= \sum_j \pi_j \mathbb{E}_{h \sim Q_j} \mathbb{E}_{\mathbf{Z} \sim \mathbf{D}_m} \hat{R}(h, \mathbf{Z}^{(j)}) \\
&= \sum_j \pi_j \mathbb{E}_{h \sim Q_j} \mathbb{E}_{\mathbf{Z}^{(j)} \sim \mathbf{D}_m^{(j)}} \hat{R}(h, \mathbf{Z}^{(j)}) \\
&= \sum_j \pi_j \mathbb{E}_{h \sim Q_n^j} R(h) \\
&= \mathbb{E}_{h \sim \pi_1 Q_n^1 + \ldots + \pi_j Q_n^j} R(h) = \mathbb{E}_{h \sim Q^{\boldsymbol{\pi}}} R(h).
\end{aligned}$$

□

Remark 4. Since the prior $\mathbf{P}_n$ and the posterior $\mathbf{Q}_n$ enter into play in this proposition and Theorem 2 through their marginals only, these results advocate for the following learning scheme. Given a cover and a (possibly factorized) prior $\mathbf{P}_n$, look for a factorized posterior $\mathbf{Q}_n = \otimes_{j=1}^n Q_j$ such that each Q_j independently minimizes the usual IID PAC-Bayes bound given in Theorem 1 on each $\mathbf{Z}^{(j)}$. Then make predictions according to the Gibbs classifier defined with respect to $Q^{\boldsymbol{\pi}} = \sum_j \pi_j Q_j$.

The following theorem gives a result that can be readily used without choosing a specific cover.

Theorem 3 (Chromatic PAC-Bayes Bound (II))**.** $\forall m$*,* $\forall \mathbf{D}_m$*,* $\forall \mathcal{H}$*,* $\forall \delta \in (0,1]$*,* $\forall P$*, with probability at least* $1-\delta$ *over the random draw of* $\mathbf{Z} \sim \mathbf{D}_m$*, the following holds*

$$\forall Q, kl(\hat{e}_Q || e_Q) \leq \frac{\chi^*}{m}\left[KL(Q||P) + \ln \frac{m+\chi^*}{\delta\chi^*}\right], \tag{4}$$

where χ^* *is the fractional chromatic number of* $\Gamma(\mathbf{D}_m)$*, and where* $\hat{e}_Q$ *and* e_Q *are defined as in (2).*

Proof. This theorem is just a particular case of Theorem 2. Let us assume that $\mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n \in \textsc{Pefc}(\mathbf{D}_m)$ such that $\omega(C) = \chi^*(\Gamma(\mathbf{D}_m))$, $\mathbf{P}_n = \otimes_{j=1}^n P = P^n$ and $\mathbf{Q}_n = \otimes_{j=1}^n Q = Q^n$, with P and Q distributions over $\mathcal{H}$.

For the right-hand side of (4), it directly comes that

$$\sum_j \alpha_j \mathrm{KL}(Q_n^j || P_n^j) = \sum_j \alpha_j \mathrm{KL}(Q||P) = \mathrm{KL}(Q||P).$$

As for the left-hand side of (4), it suffices to show that $\bar{e}_{\mathbf{Q}_n} = \hat{e}_Q$:

$$\begin{aligned}\bar{e}_{\mathbf{Q}_n} &= \sum_j \pi_j \mathbb{E}_{h\sim Q_n^j} \hat{R}(h, \mathbf{Z}^{(j)}) \\ &= \sum_j \pi_j \mathbb{E}_{h\sim Q} \hat{R}(h, \mathbf{Z}^{(j)}) \\ &= \frac{1}{m}\sum_j \omega_j |C_j| \mathbb{E}_{h\sim Q} \frac{1}{|C_j|}\sum_k r(h, Z_k) \\ &= \mathbb{E}_{h\sim Q}\frac{1}{m}\sum_j \omega_j \sum_k r(h, Z_k) \\ &= \mathbb{E}_{h\sim Q}\frac{1}{m}\sum_i r(h, Z_i) = \mathbb{E}_{h\sim Q}\hat{R}(h,\mathbf{Z}) = \hat{e}_Q.\end{aligned}$$

□

Remark 5. This theorem says that even in the case of non IID data, a PAC-Bayes bound very similar to the IID PAC-Bayes bound (1) can be stated, with a worsening (since $\chi^* \geq 1$) proportional to χ^*, i.e proportional to the amount of dependencies that exist in the data under consideration. In addition, the new PAC-Bayes bounds is valid with any priors and posteriors, without the need for these distributions nor their marginals to depend on the structure of the dependency graph, or, in other words, on the chosen cover (as is the case with the more general Theorem 2).

Remark 6. We note that among all elements of PEFC$(\mathbf{D}_m)$, χ^* is the best constant achievable in terms of the tightness of the bound. Indeed, the function $f_{m,\delta}(\omega) = \omega \ln \frac{m+\omega}{\delta\omega}$ is nondecreasing for all $m \in \mathbb{N}$ and $\delta \in (0,1]$, as indicated by the sign of the derivative $f'_{m,\delta}$:

$$\begin{aligned}f'_{m,\delta}(\omega) &= -\ln\frac{\delta\omega}{m+\omega} + \frac{\omega}{m+\omega} - 1 \\ &\geq -\ln\frac{\omega}{m+\omega} + \frac{\omega}{m+\omega} - 1 \\ &\geq -\frac{\omega}{m+\omega} + 1 + \frac{\omega}{m+\omega} - 1 = 0\end{aligned}$$

where we have used the well-known inequality $\ln x \leq x - 1$. Since χ^* is the smallest chromatic weight, this actually is the weight that gives the tightest bound.

3 Proof of Theorem 2

A proof in three steps, following the lines of the proofs given in (Seeger, 2002) and (Langford, 2005) for the IID PAC-Bayes bound, can be provided for Theorem 2.

Lemma 2. *$\forall m$, $\forall \mathbf{D}_m$, $\forall \delta \in (0,1]$, $\forall \mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n$, $\forall \mathbf{P}_n$, with probability at least $1-\delta$ over the random draw of $\mathbf{Z} \sim \mathbf{D}_m$, the following holds*

$$\mathbb{E}_{\mathbf{h}\sim\mathbf{P}_n} \sum_{j=1}^n \alpha_j e^{|C_j| kl(\hat{R}(h_j,\mathbf{Z}^{(j)})||R(h_j))} \leq \frac{m+\omega}{\delta\omega}, \quad (5)$$

where ω stands for $\omega(\mathbf{C})$.

Proof. We first observe the following:

$$\begin{aligned}&\mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m}\sum_j \alpha_j e^{|C_j|\mathrm{kl}(\hat{R}(h_j,\mathbf{Z}^{(j)})||R(h_j))} \\ &= \sum_j \alpha_j \mathbb{E}_{\mathbf{Z}^{(j)}\sim\mathbf{D}_m^{(j)}} e^{|C_j|\mathrm{kl}(\hat{R}(h,\mathbf{Z}^{(j)})||R(h))} \\ &\leq \sum_j \alpha_j(|C_j|+1) \qquad \text{(Lemma 5, Appendix)} \\ &= \frac{1}{\omega}\sum_j \omega_j(|C_j|+1) = \frac{m+\omega}{\omega},\end{aligned}$$

where using Lemma 5 is made possible by the fact that $\mathbf{Z}^{(j)}$ are IID. Therefore,

$$\mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m}\mathbb{E}_{\mathbf{h}\sim\mathbf{P}_n}\sum_{j=1}^n \alpha_j e^{|C_j|\mathrm{kl}(\hat{R}(h_j,\mathbf{Z}^{(j)})||R(h_j))} \leq \frac{m+\omega}{\omega}.$$

Applying Markov's inequality (Theorem 7, Appendix) to the random variable $\mathbb{E}_{\mathbf{h}\sim\mathbf{P}_n}\sum_j \alpha_j e^{|C_j|\mathrm{kl}(\hat{R}(h_j,\mathbf{Z}^{(j)})||R(h_j))}$ gives the desired result. □

Lemma 3. *$\forall m$, $\forall \mathbf{D}_m$, $\forall \mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n$, $\forall \mathbf{P}_n$, $\forall \mathbf{Q}_n$, with probability at least $1-\delta$ over the random draw of $\mathbf{Z} \sim \mathbf{D}_m$, the following holds*

$$\frac{m}{\omega}\sum_{j=1}^n \pi_j \mathbb{E}_{h\sim Q_n^j} kl(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)) \quad (6)$$
$$\leq \sum_{j=1}^n \alpha_j KL(Q_n^j||P_n^j) + \ln\frac{m+\omega}{\delta\omega}.$$

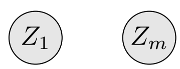
Proof. It suffices to use Jensen's inequality with $\ln$ and the fact that $\mathbb{E}_{X\sim P} f(X) = \mathbb{E}_{X\sim Q}\frac{P(X)}{Q(X)}f(X)$, for all f, P, Q. Therefore, $\forall \mathbf{Q}_n$:

$$\begin{aligned}&\ln\mathbb{E}_{\mathbf{h}\sim\mathbf{P}_n}\sum_j \alpha_j e^{|C_j|\mathrm{kl}\left(\hat{R}(h_j,\mathbf{Z}^{(j)})||R(h_j)\right)} \\ &= \ln\sum_j \alpha_j \mathbb{E}_{h\sim P_n^j} e^{|C_j|\mathrm{kl}\left(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)\right)} \\ &= \ln\sum_j \alpha_j \mathbb{E}_{h\sim Q_n^j}\frac{P_n^j(h)}{Q_n^j(h)} e^{|C_j|\mathrm{kl}\left(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)\right)} \\ &\geq \sum_j \alpha_j \mathbb{E}_{h\sim Q_n^j}\ln\left[\frac{P_n^j(h)}{Q_n^j(h)} e^{|C_j|\mathrm{kl}\left(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)\right)}\right] \\ &= -\sum_j \alpha_j \mathrm{KL}(Q_n^j||P_n^j) \\ &\quad + \sum_j \alpha_j |C_j| \mathbb{E}_{h\sim Q_n^j}\mathrm{kl}\left(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)\right) \\ &= -\sum_j \alpha_j \mathrm{KL}(Q_n^j||P_n^j) \\ &\quad + \frac{m}{\omega}\sum_j \pi_j \mathbb{E}_{h\sim Q_n^j}\mathrm{kl}\left(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)\right).\end{aligned}$$

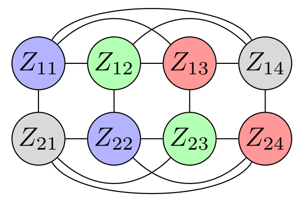
Lemma 2 then gives the result. □

Lemma 4. *$\forall m$, $\forall \mathbf{D}_m$, $\forall \mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n$, $\forall \mathbf{Q}_n$,, the following holds*

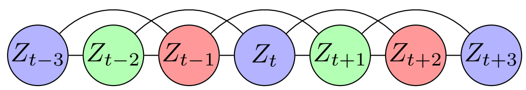
$$\frac{m}{\omega}\sum_{j=1}^n \pi_j \mathbb{E}_{h\sim Q_n^j} kl(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)) \geq kl(\bar{e}_Q||e_Q).$$



(a) IID data



(b) Bipartite Ranking data



(c) Sliding window data

Figure 1: Dependency graphs for the different settings described in section 4. Nodes of the same color are part of the same cover element; henceforth, they are independent. (a) When the data are IID , the dependency graph is disconnected and the fractional number is $\chi^* = 1$; (b) a dependency graph obtained for bipartite ranking from a sample containing 4 positive instances and 2 negative instances: $\chi^* = 4$; (c) a dependency graph obtained with the technique of sliding windows for sequence data, for a window parameter $r = 1$ (see text for details): $\chi^* = 2r + 1$.

Proof. This simply comes from the application of Theorem 6 given in Appendix. This lemma, in combination with Lemma 3, closes the proof of Theorem 2. $\square$

4 Examples

In this section, we give instances of the bound given in Theorem 3 for different settings.

4.1 IID case

In the IID case, the training sample is $\mathbf{Z} = \{(X_i, Y_i)\}_{i=1}^m$ distributed according to $\mathbf{D}_m = D^m$ and the fractional chromatic number of $\Gamma(\mathbf{D}_m)$ is $\chi^* = 1$. Plugging in this value of χ^* in the bound of Theorem 3 gives the usual PAC-Bayes bound recalled in Theorem 1. This emphasizes the fact that the standard PAC-Bayes bound is a special case of our more general results.

4.2 Bipartite Ranking

Let $\overline{D}$ be a distribution over $\overline{\mathcal{X}} \times \overline{\mathcal{Y}}$ and $\overline{D}_{+1}$ ($\overline{D}_{-1}$) be the class conditional distribution $\overline{D}_{X|Y=+1}$ ($\overline{D}_{X|Y=-1}$) with respect to $\overline{D}$. In the bipartite ranking problem (see, e.g. (Agarwal et al., 2005)), one tries to control the misranking risk, defined for $f \in \mathbb{R}^{\mathcal{X}}$ by

$$R^{\text{rank}}(f) = \mathbb{P}_{\substack{\overline{X}^+ \sim \overline{D}_{+1} \\ \overline{X}^- \sim \overline{D}_{-1}}} (f(\overline{X}^+) \leq f(\overline{X}^-)). \tag{7}$$

f can be interpreted as a scoring function. Given an IID sample $\mathbf{S} = \{(\overline{X}_i, \overline{Y}_i)\}_{i=1}^{\ell}$ distributed according to $\overline{\mathbf{D}}_\ell = \overline{D}^\ell$, a usual strategy to minimize (7) is to minimize (a possibly regularized form of)

$$\hat{R}^{\text{rank}}(f, \mathbf{S}) = \frac{1}{\ell^+\ell^-} \sum_{\substack{i:\overline{Y}_i=+1 \\ j:\overline{Y}_j=-1}} r(f, (\overline{X}_i, \overline{X}_j)), \tag{8}$$

where $r(f, (\overline{X}_i, \overline{X}_j)) = \mathbb{I}_{f(\overline{X}_i) \leq f(\overline{X}_j)}$ and ℓ^+ (ℓ^-) is the number of positive (negative) data in $\mathbf{S}$. This empirical risk, which is closely related to the Area under the ROC curve, or AUC[1] (Agarwal et al., 2005; Cortes & Mohri, 2004), estimates the fraction of pairs $(\overline{X}_i, \overline{X}_j)$ that are ranked incorrectly (given that $\overline{Y}_i = +1$ and $\overline{Y}_j = -1$) and is an unbiased estimator of $R^{\text{rank}}(h)$. The entailed problem can be seen as that of learning a classifier from a training set of the form $\mathbf{Z} = \{Z_{ij}\}_{ij} = \{(X_{ij} = (\overline{X}_i, \overline{X}_j), 1)\}_{ij}$. This reveals the non-IID nature of the training data, as Z_{ij} depends on $\{Z_{pq} : p = i \text{ or } q = j\}$ (see Figure 1).

Using Theorem 3, we have the following result:

Theorem 4. *$\forall \ell$, $\forall \overline{D}$ over $\overline{\mathcal{X}} \times \overline{\mathcal{Y}}$, $\forall \overline{\mathcal{H}} \subseteq \mathbb{R}^{\overline{\mathcal{X}}}$, $\forall \delta \in (0, 1]$, $\forall \overline{P}$ over $\overline{\mathcal{H}}$, with probability at least $1 - \delta$ over the random draw of $\mathbf{S} \sim \overline{D}^\ell$, the following holds*

$$\forall \overline{Q} \text{ over } \overline{\mathcal{H}}, kl(\hat{e}_{\overline{Q}}^{rank} || e_{\overline{Q}}^{rank}) \leq \frac{1}{\ell_{\min}} \left[KL(\overline{Q}||\overline{P}) + \ln \frac{\ell_{\min} + 1}{\delta} \right], \tag{9}$$

where $\ell_{\min} = \min(\ell^+, \ell^-)$, and $\hat{e}_{\overline{Q}}^{rank}$ and $e_{\overline{Q}}^{rank}$ are the Gibbs ranking error counterparts of (2) based on (7) and (8), respectively.

Proof. The proof works in three parts and borrows ideas from (Agarwal et al., 2005). The first two parts are necessary to deal with the fact that the dependency graph of $\mathbf{Z}$, as implied by $\mathbf{S}$, does not have a deterministic structure.

Conditioning on $\mathbf{Y} = \mathbf{y}$. Let $\mathbf{y} \in \{-1, +1\}^\ell$ be a fixed vector and $\ell_{\mathbf{y}}^+$ and $\ell_{\mathbf{y}}^-$ the number of positive and negative labels, respectively. We define the distribution $\overline{\mathbf{D}}_{\mathbf{y}}$ as $\overline{\mathbf{D}}_{\mathbf{y}} = \otimes_{i=1}^{\ell} \overline{D}_{y_i}$; this is a distribution on $\overline{\mathcal{X}}^\ell$. With a slight abuse of notation, $\overline{\mathbf{D}}_{\mathbf{y}}$ will also be used to denote the distribution over $(\overline{\mathcal{X}} \times \overline{\mathcal{Y}})^\ell$ of samples $\mathbf{S} = \{(\overline{X}_i, y_i)\}_{i=1}^{\ell}$ such that the sequence $\{\overline{X}_i\}_{i=1}^{\ell}$ is distributed according to $\overline{\mathbf{D}}_{\mathbf{y}}$. It is straightforward to check that, $\forall f \in \overline{\mathcal{H}}$, $\mathbb{E}_{\mathbf{S} \sim \overline{\mathbf{D}}_{\mathbf{y}}} \hat{R}^{\text{rank}}(f, \mathbf{S}) = R^{\text{rank}}(f)$ (cf. Equation (7)).

Given $\mathbf{S}$, defining the random variable Z_{ij} as $Z_{ij} = ((X_i, X_j), 1)$, $\mathbf{Z} = \{Z_{ij}\}_{i:y_i=1, j:y_j=-1}$ is a sample of identically distributed variables, each with distribution $D_{\pm 1} = \overline{D}_{+1} \otimes \overline{D}_{-1} \otimes \mathbf{1}$ over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} = \overline{\mathcal{X}} \times \overline{\mathcal{X}}$, $\mathcal{Y} = \{-1, +1\}$ and $\mathbf{1}$ is the distribution that produces 1 with probability 1.

[1]It is actually 1-AUC.

Letting $m = \ell_{\mathbf{y}}^+\ell_{\mathbf{y}}^-$ we denote by $\mathbf{D}_{\mathbf{y},m}$ the distribution of the training sample $\mathbf{Z}$, within which interdependencies exist, as shown on Figure 1. Theorem 2 can thus be directly applied to classifiers trained on $\mathbf{Z}$, the structure of $\Gamma(\mathbf{D}_{\mathbf{y},m})$ and its corresponding fractional chromatic number $\chi_{\mathbf{y}}^*$ being completely determined by $\mathbf{y}$. Letting $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, $\forall\delta \in (0,1]$, $\forall P$ over $\mathcal{H}$, with probability at least $1-\delta$ over the random draw of $\mathbf{Z} \sim \mathbf{D}_{\mathbf{y},m}$,

$$\forall Q \text{ over } \mathcal{H}, \mathrm{kl}(\hat{e}_Q||e_Q) \leq \frac{\chi_{\mathbf{y}}^*}{m}\left[\mathrm{KL}(Q||P) + \ln\frac{m+\chi_{\mathbf{y}}^*}{\delta\chi_{\mathbf{y}}^*}\right].$$

Given $f \in \overline{\mathcal{H}}$, it is straightforward to see that for $h_f \in \mathcal{Y}^{\mathcal{X}}$ defined as $h_f((X,X')) = \mathrm{sign}(f(X) - f(X'))$, with $\mathrm{sign}(x) = +1$ if $x > 0$ and -1 otherwise, $\hat{R}(h_f, \mathbf{Z}) = \hat{R}^{\text{rank}}(f, \mathbf{S})$ and $\mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_{\mathbf{y},m}}\hat{R}(h_f,\mathbf{Z}) = \mathbb{E}_{\mathbf{S}\sim\overline{\mathbf{D}}_{\mathbf{y}}}\hat{R}^{\text{rank}}(f,\mathbf{S}) = R^{\text{rank}}(f)$. Hence, $\forall\delta \in (0,1]$, $\forall\overline{P}$ over $\overline{\mathcal{H}}$, with probability at least $1-\delta$ over the random draw of $\mathbf{S} \sim \overline{\mathbf{D}}_{\mathbf{y}}$,

$$\forall\overline{Q}, \mathrm{kl}(\hat{e}_{\overline{Q}}^{\text{rank}}||e_{\overline{Q}}^{\text{rank}}) \leq \frac{\chi_{\mathbf{y}}^*}{m}\left[\mathrm{KL}(\overline{Q}||\overline{P}) + \ln\frac{m+\chi_{\mathbf{y}}^*}{\delta\chi_{\mathbf{y}}^*}\right]. \quad (10)$$

Integrating over Y. As proposed in (Agarwal et al., 2005), let us call $\Phi(\overline{P}, \mathbf{S}, \delta)$ the event (10); we just stated that $\forall\mathbf{y} \in \{-1,+1\}^\ell$, $\forall\overline{P}$, $\forall\delta \in (0,1]$, $\mathbb{P}_{\mathbf{S}\sim\overline{\mathbf{D}}_{\mathbf{y}}}(\Phi(\overline{P},\mathbf{S},\delta)) \geq 1-\delta$. Then, $\forall\overline{P}, \forall\delta \in (0,1]$,

$$\begin{aligned}
\mathbb{P}_{\mathbf{S}\sim\overline{\mathbf{D}}_\ell}(\Phi(\overline{P},\mathbf{S},\delta)) &= \mathbb{E}_{\mathbf{Y}}[\mathbb{E}_{\mathbf{S}\sim\overline{\mathbf{D}}_{\mathbf{Y}}}\mathbb{I}_{\Phi(\overline{P},\mathbf{S},\delta)}] \\
&= \sum_{\mathbf{y}} \mathbb{E}_{\mathbf{S}\sim\overline{\mathbf{D}}_{\mathbf{y}}}\mathbb{I}_{\Phi(\overline{P},\mathbf{S},\delta)}\mathbb{P}(\mathbf{Y}=\mathbf{y}) \\
&= \sum_{\mathbf{y}} \mathbb{P}_{\mathbf{S}\sim\overline{\mathbf{D}}_{\mathbf{y}}}(\Phi(\overline{P},\mathbf{S},\delta))\mathbb{P}(\mathbf{Y}=\mathbf{y}) \\
&\geq \sum_{\mathbf{y}} (1-\delta)\mathbb{P}(\mathbf{Y}=\mathbf{y}) = 1-\delta.
\end{aligned}$$

Hence, $\forall\delta \in (0,1]$, $\forall\overline{P}$ over $\overline{\mathcal{H}}$, with probability at least $1-\delta$ over the random draw of $\mathbf{S} \sim \overline{\mathbf{D}}_\ell$,

$$\forall\overline{Q}, \mathrm{kl}(\hat{e}_{\overline{Q}}^{\text{rank}}||e_{\overline{Q}}^{\text{rank}}) \leq \frac{\chi_{\mathbf{S}}^*}{m_{\mathbf{S}}}\left[\mathrm{KL}(\overline{Q}||\overline{P}) + \ln\frac{m_{\mathbf{S}}+\chi_{\mathbf{S}}^*}{\delta\chi_{\mathbf{S}}^*}\right]. \quad (11)$$

where $\chi_{\mathbf{S}}^*$ is the fractional chromatic number of the graph $\Gamma(\mathbf{Z})$, with $\mathbf{Z}$ defined from $\mathbf{S}$ as in the first part of the proof (taking into account the observed labels in $\mathbf{S}$); here $m_{\mathbf{S}} = \ell^+\ell^-$, where ℓ^+ (ℓ^-) is the number of positive (negative) data in $\mathbf{S}$.

Computing the Fractional Chromatic Number. In order to finish the proof, it suffices to observe that, for $\mathbf{Z} = \{Z_{ij}\}_{ij}$, letting $\ell_{\max} = \max(\ell^+,\ell^-)$, the fractional chromatic number of $\Gamma(\mathbf{Z})$ is $\chi^* = \ell_{\max}$.

Indeed, the clique number of $\Gamma(\mathbf{Z})$ is $\ell_{\max}$ as for all $i = 1,\ldots,\ell^+$ ($j = 1,\ldots,\ell^-$), $\{Z_{ij} : j = 1,\ldots,\ell^-\}$ ($\{Z_{ij} : i = 1,\ldots,\ell^+\}$) defines a clique of order ℓ^- (ℓ^+) in $\Gamma(\mathbf{Z})$. Thus, from Property 1: $\chi \geq \chi^* \geq \ell_{\max}$.

A proper exact cover $\mathbf{C} = \{C_k\}_{k=1}^{\ell_{\max}}$ of $\Gamma(\mathbf{Z})$ can be constructed as follows[2]. Suppose that $\ell_{\max} = \ell^+$, then $C_k = \{Z_{i\sigma_k(i)} : i = 1,\ldots,\ell^-\}$, with

$$\sigma_k(i) = (i + k - 2 \mod \ell^+) + 1,$$

is an independent set: no two variables Z_{ij} and Z_{pq} in C_k are such that $i = p$ or $j = q$. In addition, it is straightforward to check that $\mathbf{C}$ is indeed a cover of $\Gamma(\mathbf{Z})$. This cover is of size $\ell^+ = \ell_{\max}$, which means that it achieves the minimal possible weight over proper exact (fractional) covers since $\chi^* \geq \ell_{\max}$. Hence, $\chi^* = \chi = \ell_{\max}(= c(\Gamma))$. Plugging in this value of χ^* in (11), and noting that $m_{\mathbf{S}} = \ell_{\max}\ell_{\min}$ with $\ell_{\min} = \min(\ell^+,\ell^-)$, closes the proof. □

As proposed by (Langford, 2005), the PAC-Bayes bound of Theorem 4 can be specialized to the case where $\overline{\mathcal{H}} = \{f : f(x) = w\cdot x, w \in \overline{\mathcal{X}}\}$. In this situation, for $f \in \overline{\mathcal{H}}$, $h_f((X,X')) = \mathrm{sign}(f(X)-f(X')) = \mathrm{sign}(w\cdot(X-X'))$ is simply a linear classifier (the following results therefore carries over to the use of kernel classifiers). Hence, assuming an isotropic Gaussian prior $P = \mathcal{N}(0,I)$ and a family of posteriors $Q_{w,\mu}$ parameterized by $w \in \overline{\mathcal{X}}$ and $\mu > 0$ such that $Q_{w,\mu}$ is $\mathcal{N}(\mu,1)$ in the direction w and $\mathcal{N}(0,1)$ in all perpendicular directions, we arrive at the following theorem (of which we omit the proof):

Theorem 5. *$\forall\ell, \forall\overline{D}$ over $\overline{\mathcal{X}}\times\overline{\mathcal{Y}}$, $\forall\delta \in (0,1]$, the following holds with prob. at least $1-\delta$ over the draw of $\mathbf{S} \sim \overline{D}^\ell$:*

$$\forall w, \mu > 0, kl(\hat{R}_{Q_{w,\mu}}^{rank}||R_{Q_{w,\mu}}^{rank}) \leq \frac{1}{\ell_{\min}}\left[\frac{\mu^2}{2} + \ln\frac{\ell_{\min}+1}{\delta}\right].$$

The bounds given in Theorem 4 and Theorem 5 are very similar to what we would get if applying IID PAC-Bayes bound to one (independent) element C_j of a minimal cover (i.e. its weight equals the fractional chromatic number) $\mathbf{C} = \{C_j\}_{j=1}^n$ such as the one we have constructed in the proof of Theorem 4. This would imply the empirical error $\hat{e}_{\overline{Q}}^{\text{rank}}$ to be computed on only one specific C_j and not all the C_j's simultaneously, as is the case for the new results. It turns out that, for proper exact fractional covers $\mathbf{C} = \{(C_j,\omega)\}_{j=1}^n$ with elements C_j having the same size, it is better, in terms of absolute moments of the empirical error, to assess it on the whole dataset, rather than on only one C_j. The following proposition formalizes this.

Proposition 2. *$\forall m$, $\forall\mathbf{D}_m$, $\forall\mathcal{H}$, $\forall\mathbf{C} = \{(C_j,\omega_j\}_{j=1}^n \in$ PEFC$(\mathbf{D}_m)$, $\forall Q$, $\forall r \in \mathbb{N}, r \geq 1$, if $|C_1| = \ldots = |C_n|$ then*

$$\mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m}|\hat{e}_Q - e_Q|^r \leq \mathbb{E}_{\mathbf{Z}^{(j)}\sim\mathbf{D}_m^{(j)}}|\hat{e}_Q^{(j)} - e_Q|^r, \forall j \in \{1,\ldots n\},$$

where $\hat{e}_Q^{(j)} = \mathbb{E}_{h\sim Q}\hat{R}(h,\mathbf{Z}^{(j)})$.

[2] Note that the cover defined here considers elements C_k containing random variables themselves instead of their indices. This abuse of notation is made for sake of readability.

Proof. It suffices to use the convexity of $|\cdot|^r$ for $r \geq 1$ and the linearity of $\mathbb{E}$. Using notation of section 2, we have, for $\mathbf{Z} \sim \mathbf{D}_m$:

$$\begin{aligned}|\hat{e}_Q - e_Q|^r &= |\sum_j \pi_j \mathbb{E}_{h\sim Q}(\hat{R}(h, \mathbf{Z}^{(j)}) - R(h))|^r \\ &\leq \sum_j \pi_j |\mathbb{E}_{h\sim Q}(\hat{R}(h, \mathbf{Z}^{(j)}) - R(h))|^r \\ &= \sum_j \pi_j |\hat{e}_Q^{(j)} - e_Q|^r.\end{aligned}$$

Taking the expectation of both sides with respect to $\mathbf{Z}$ and noting that the random variables $|\hat{e}_Q^{(j)} - e_Q|^r$, have the same distribution, gives the result. □

4.3 Sliding Windows for Sequence Data

There are many situations, such as in bioinformatics, where a classifier must be learned from a training sample $\mathbf{S} = \{(\overline{X}_t, \overline{Y}_t)\}_{t=1}^T \in (\overline{\mathcal{X}} \times \overline{\mathcal{Y}})^T$ where it is known that there is a sequential dependence between the X_t's. A typical approach to tackle the problem of learning from such data is the following: in order to predict $\overline{Y}_t$, information from a *window* $\{\overline{X}_{t+\tau}\}_{\tau=-r}^r$ of $2r+1$ data centered on $\overline{X}_t$ is considered, r being set according to some prior knowledge or after a cross-validation process. This problem can be cast in another classification problem using a training sample $\mathbf{Z} = \{Z_t\}_{t=1}^T$, with $Z_t = ((\overline{X}_{t-r}, \ldots, \overline{X}_t, \ldots, \overline{X}_{t+r}), \overline{Y}_t)$, with special care taken for $t \leq r+1$ and $t > T - r$. Considering that $\overline{\mathcal{Y}} = \{-1, +1\}$, the input space and output space to be considered are therefore $\mathcal{X} = \overline{\mathcal{X}}^{2r+1}$ and $\mathcal{Y} = \overline{\mathcal{Y}}$; the product space is $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. As for the bipartite ranking problem, we end up with a learning problem from non-IID data, $\mathbf{Z}$ having a dependency graph $\Gamma(\mathbf{Z})$ as the one depicted on Figure 1.

It is easy to see that the clique number of $\Gamma(\mathbf{Z})$ is $2r+1$. Besides, one can construct a proper exact cover $\mathbf{C} = \{C_j\}_{j=1}^{2r+1}$ of minimal size/weight by taking $C_j = \{Z_{j+p(2r+1)} : p = 0, \ldots, \lfloor \frac{T-j}{2r+1} \rfloor\}$, for $j = 1, \ldots, 2r+1$ – we make the implicit and reasonable assumption that $T > 2r+1$. This cover is proper and has size $2r+1$. Invoking Property 1 gives that $\chi = \chi^* = 2r+1$.

It is therefore easy to get a new PAC-Bayes theorem for the case of windowed prediction, by replacing χ^* by $2r+1$ and m by T in the bound (4) of Theorem 3. We do not state it explicitly for sake of conciseness.

5 Conclusion

In this work, we propose the first PAC-Bayes bounds applying for classifiers trained on non-IID data. The derivation of these results rely on the use of fractional covers of graphs, convexity and standard tools from probability theory. The results that we provide are very general and can easily be instantiated for specific learning settings such as bipartite ranking and windowed prediction for sequence data.

This work gives rise to many interesting questions. First, it seems that using a fractional cover to decompose the non-IID training data into sets of IID data and then tightening the bound through the use of the chromatic number is some form of variational relaxation as often encountered in the context of inference in graphical models, the graphical model under consideration in this work being one that encodes the dependencies in $\mathbf{D}_m$. It might be interesting to make this connection clearer to see if, for instance, tighter and still general bounds can be obtained with more appropriate variational relaxations than the one incurred by the use of fractional covers.

Besides, Theorem 2 advocates for the learning algorithm described in Remark 4. It would be interesting to see how such a learning algorithm based on possibly multiple priors/multiple posteriors could perform empirically and how tight the proposed bound could be.

On another empirical side, we are planning to run intensive numerical simulations on bipartite ranking problems to see how accurate the bound of Theorem 5 can be: we expect the results to be of good quality, because of the resemblance of the bound of the theorem with the IID PAC-Bayes theorem for margin classifiers, which has proven to be rather accurate (Langford, 2005). Likewise, it would be interesting to see how the possibly more accurate PAC-Bayes bound for large margin classifiers proposed by (Langford & Shawe-taylor, 2002), which should translate to the case of bipartite ranking as well, performs empirically.

It also remains the question as to what kind of strategies to learn the prior(s) could be implemented to render the bound of Theorem 2 the tightest possible. This is one of the most stimulating question as performing such prior learning makes it possible to obtain very accurate generalization bound, as evidenced by (Ambroladze et al., 2007).

Finally, assuming the data are identically distributed might be too strong an assumption. This brings up the question on whether it is possible to derive the same kind of results as those provided here in the case where the variables do not have the same marginals: we have recently obtained a positive answer on deriving such a bound (Ralaivola, 2009), by directly leveraging a concentration inequality given in (Janson, 2004). We are also currently investigating how PAC-Bayes bounds could be derived for a different setting that gives rise to non-IID data, namely mixing processes.

Acknowledgment

This work is partially supported by the IST Program of the EC, under the FP7 Pascal 2 Network of Excellence, ICT-216886-NOE.

Appendix

Lemma 5. *Let D be a distribution over $\mathcal{Z}$.*

$$\forall h \in \mathcal{H}, \mathbb{E}_{\mathbf{Z}\sim D^m} e^{mkl\left(\hat{R}(h,\mathbf{Z})||R(h)\right)} \leq m+1.$$

Proof. Let $h \in \mathcal{H}$. For $\mathbf{z} \in \mathcal{Z}^m$, we let $q(\mathbf{z}) = \hat{R}(h, \mathbf{z})$; we also let $p = R(h)$. Note that since $\mathbf{Z}$ is i.i.d, $mq(\mathbf{Z})$ is binomial with parameters m and p (recall that $r(h, Z)$ takes the values 0 and 1 upon correct and erroneous classification of Z by h, respectively).

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Z}\sim D^m} e^{m\mathrm{kl}(q(\mathbf{Z})||p)} \\
&\quad = \sum_{\mathbf{z}\in\mathcal{Z}^m} e^{m\mathrm{kl}(q(\mathbf{z})||p)} \mathbb{P}_{\mathbf{Z}\sim D^m}(\mathbf{Z}=\mathbf{z}) \\
&\quad = \sum_{0\leq k\leq m} e^{m\mathrm{kl}\left(\frac{k}{m}||p\right)} \mathbb{P}_{\mathbf{Z}\sim D^m}(mq(\mathbf{Z})=k) \\
&\quad = \sum_{0\leq k\leq m} \binom{m}{k} e^{m\mathrm{kl}\left(\frac{k}{m}||p\right)} p^k(1-p)^{m-k} \\
&\quad = \sum_{0\leq k\leq m} \binom{m}{k} e^{m\left(\frac{k}{m}\ln\frac{k}{m}+(1-\frac{k}{m})\ln(1-\frac{k}{m})\right)} \\
&\quad = \sum_{0\leq k\leq m} \binom{m}{k} \left(\frac{k}{m}\right)^k \left(1-\frac{k}{m}\right)^{m-k}.
\end{aligned}$$

However, it is obvious that, from the definition of the binomial distribution,

$$\forall m \in \mathbb{N}, \forall k \in [0, m], \forall t \in [0, 1], \binom{m}{k} t^k(1-t)^{m-k} \leq 1.$$

This is obviously the case for $t = \frac{k}{m}$, which gives

$$\sum_{0\leq k\leq m} \binom{m}{k} \left(\frac{k}{m}\right)^k \left(1-\frac{k}{m}\right)^{m-k} \leq \sum_{0\leq k\leq m} 1 = m+1.$$

$\square$

Theorem 6 (Jensen's inequality)**.** *Let $f \in \mathbb{R}^{\mathcal{X}}$ be a convex function. For all probability distribution P on $\mathcal{X}$:*

$$f(\mathbb{E}_{X\sim P}X) \leq \mathbb{E}_{X\sim P}f(X).$$

Proof. Directly comes by induction on the definition of a convex function. $\square$

Theorem 7 (Markov's Inequality)**.** *Let X be a positive random variable on $\mathbb{R}$, such that $\mathbb{E}X < \infty$.*

$$\forall t \in \mathbb{R}, \mathbb{P}_X\left\{X \geq \frac{\mathbb{E}X}{t}\right\} \leq \frac{1}{t}.$$

Consequently: $\forall M \geq \mathbb{E}X, \forall t \in \mathbb{R}, \mathbb{P}_X\left\{X \geq \frac{M}{t}\right\} \leq \frac{1}{t}$.

Proof. In almost all textbooks on probability. $\square$

Lemma 6. $\forall p, q, r, s \in [0, 1], \forall \alpha \in [0, 1]$,

$$\begin{aligned}
&kl(\alpha p + (1-\alpha)q||\alpha r + (1-\alpha)s) \\
&\qquad\qquad \leq \alpha kl(p||r) + (1-\alpha)kl(q||s).
\end{aligned}$$

Proof. It suffices to see that $f \in \mathbb{R}^{[0,1]^2}, f(\mathbf{v} = [p\ q]) = \mathrm{kl}(q||p)$ is convex over $[0, 1]^2$: the Hessian H of f is

$$H = \begin{bmatrix} \frac{q}{p^2} + \frac{1-q}{(1-p)^2} & -\frac{1}{p} - \frac{1}{1-p} \\ -\frac{1}{p} - \frac{1}{1-p} & \frac{1}{q} + \frac{1}{1-q} \end{bmatrix},$$

and, for $p, q \in [0, 1]$, $\frac{q}{p^2} + \frac{1-q}{(1-p)^2} \geq 0$ and $\det H = \frac{(p-q)^2}{q(1-q)p^2(1-p)^2} \geq 0$: $H \succeq 0$ and f is indeed convex. $\square$

References

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., & Roth, D. (2005). Generalization Bounds for the Area Under the ROC Curve. *J. of Machine Learning Research*, *6*, 393–425.

Ambroladze, A., Parrado-Hernandez, E., & Shawe-Taylor, J. (2007). Tighter PAC-Bayes Bounds. *Adv. in Neural Information Processing Systems 19* (pp. 9–16).

Cortes, C., & Mohri, M. (2004). AUC Optimization vs. Error Rate Minimization. *Adv. in Neural Information Processing Systems 16*.

Janson, S. (2004). Large Deviations for Sums of Partly Dependent Random Variables. *Random Structures Algorithms*, *24*, 234–248.

Langford, J. (2005). Tutorial on Practical Theory for Classification. *J. of Machine Learning Research*, 273–306.

Langford, J., & Shawe-taylor, J. (2002). PAC-Bayes and Margins. *Adv. in Neural Information Processing Systems 15* (pp. 439–446).

McAllester, D. (1999). Some PAC-Bayesian Theorems. *Machine Learning*, *37*, 355–363.

Ralaivola, L. (2009). PAC-Bayes Bounds and non IID Data: Variations on a Theme. In preparation.

Schreinerman, E., & Ullman, D. (1997). *Fractional graph theory: A rational approach to the theory of graphs*. Wiley Interscience Series in Discrete Mathematics.

Seeger, M. (2002). PAC-Bayesian generalization bounds for gaussian processes. *J. of Machine Learning Research*, *3*, 233–269.

Usunier, N., Amini, M.-R., & Gallinari, P. (2006). Generalization Error Bounds for Classifiers Trained with Interdependent Data. *Adv. in Neural Information Processing Systems 18* (pp. 1369–1376).